Učenje akustičkih modela govora za raspoznavanje pomoću alata HTK/Julius

Nea Dujmić

30. kolovoza 2020.

Sažetak

Cilj projekta je izrada sustava za raspoznavanje pojedinih riječi koristeći programske alate HTK i Julius. Uz učenje akustičkog modela, kroz projekt će biti opisan postupak modeliranja govora pomoću prikrivenih Markovljevih modela.

1 Uvod

Potreba za raspoznavanjem govora pojavljuje se kroz mnoge aplikacije računalnih programa. Korisnik neovisno o načinu korištenja ima potrebu većinu interakcija s računalom provoditi koristeći govor umjesto upisivanja podataka. To ne znači da je glasovno upravljanje primjenjivo u svim područjima te da postoji potreba za potpunom zamjenom, ali pojednostavljuje mnoge primjene.
Trenutno postignuta razina točnosti raspoznavanja govora limitira svakodnevnu uporabu na jednostavne primjene. Dobar pokazatelj je korištenje uređaja za glasovno upravljanje kao što su (Amazon) Echo i Google Home primarno za mijenjanje i biranje glazbe te korištenje štoperice. Glasovni pomagači koji dolaze kao dio mogućnosti koje pametni telefoni nude se najčešće koristi za biranje kontakata kod poziva.
Korištenje kod većih sustava je najčešće ograničeno zbog nepoznavanja podataka od strane korisnika, dok je taj problem puno lakše rješiv primjerice padajućim izbornikom ili automatskim dopunjavanjem kod grafičkih sučelja.

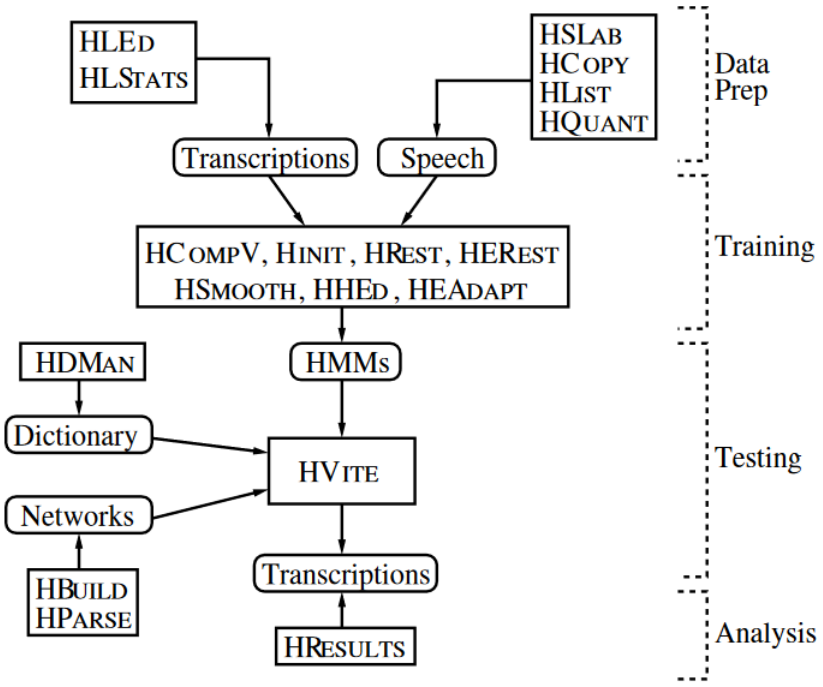
2 Korišteni alati

Za izradu akustičkog modela za raspoznavanje korišteni su računalni programi Julius te HTK (Hidden Markov Model Toolkit). Kod pripreme podataka korištene su skripte pisane u programskom jeziku Julia.

2.1 HTK

Prikriveni Markovljevi lanci implementirani su kao dio alata HTK kroz funkcije za manipulaciju podataka potrebnih pri učenju akustičkog modela.
Osim kod raspoznavanja govora, HTK svoju primjenu pronalazi i kod sinteze govora, raspoznavanja ručno pisanih znakova te kao alat kod sekvenciranja DNK. Kako se HTK zasniva na uporabi prikrivenih Markovljevih modela može se generalno koristiti za izradu modela podataka koji se ponavljaju u određenim vremenskim periodima (engl. time-series).
Korištenje HTK alata podrazumijeva dvije faze. Prva faza je procjena parametara za prikriveni

Markovljev model koristeći HTK alate za treniranje modela u kojoj koristimo govorne iskaze (engl. utterance) i pridružene im transkripcije. Druga faza je uporaba HTK alata za prepoznavanje nad transkripcijama nepoznatih govornih iskaza.



Slika 1: Dostupni HTK alati kroz faze izrade modela

2.2 Julius

Julius je program otvorenog koda koji nudi dekodiranje prije snimljenih datoteka govora kao i onih danih u stvarnom vremenu. Pruža veliku brzinu dekodiranja, dobre performanse, modularnost te dostupnost.
Može koristiti akustičke modele kreirane pomoću HTK alat, ali koristi i svoj format korištenja gramatike. Za pokretanje je potrebna konfiguracijska datoteka u kojoj uz naznačene parametre i razine naznačavamo i putanje do rječnika izgovora (engl. pronunciation dictionary), akustičnog prikrivenog Markovljevog modela (engl. acoustic HMM), datoteke koja mapira logičke foneme u fizičke (engl. HMMList to map logical phone to physical) te gramatičku datoteku automata konačnih stanja (engl. finite state automaton grammar file).
Potrebne datoteke dobivene se koristeći HTK alate te može prepoznati samo izraze koji su naznačeni u gramatičkoj datoteci (engl. grammar).

3 Prikriveni Markovljevi model

Jedan od najpopularnijih modela koji se primjenjuje nad sekvencijalnim podacima. Učestalo korištenje i popularnost su posljedice jednostavnosti modela tj. jednostavnosti određivanja (učenja) parametara, posebice ako usporedimo s određivanjem parametara kod neuralnih mreža.

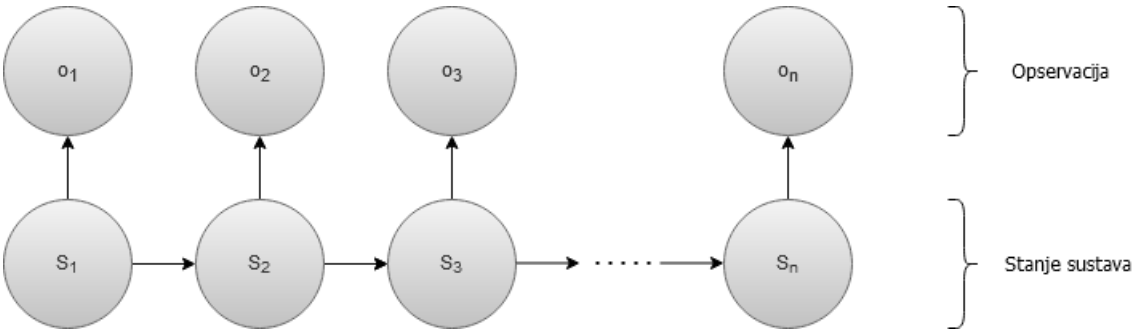
Prikriveni Markovljev model je verzija Markovljevog modela u kojoj se za modelirani sustav pretpostavlja da se ponaša kao Markovljev proces. Za razliku od Markovljevog modela za koji možemo izravno promatrati i uočavati prelaske iz jednog stanja sutava u drugo, kod prikrivenog Markovljevog modela su stanja sustava prikrivena. Umjesto izravnog uočavanja stanja sustava možemo napraviti opservaciju događaja.

Skup stanja S sadrži stanja u kojima se sustav može nalaziti, a koja istovremeno ne možemo opaziti.

- $S = \{S_1, ..., S_n\}$

Za razliku od skupa stanja S, skup O označava opservacije.

- $O = \{o_1, ..., o_m\}$



Slika 2: Prikriveni Markovljev model

Broj stanja sustava i mogućih opservacija je međusobno neovisan, a vezani su vjerojatnošću da je sustav u stanju S_x za trenutnu opservaciju o_x.
Skup π definira vjerojatnosti da određeno stanje koje sustav može poprimiti bude početno.
Vremena u kojima se bilježe opservacije bilježimo varijablom t: $1 \leq t \leq T$.

3.1 Tri glavne uporabe

3.1.1 Vjerojatnost uočene sekvence

Evaluacijski problem izračunavanja vjerojatnosti da je uočena sekvenca rezultat modela ili procjena opisuje li model dovoljno dobro uočenu sekvencu. Kod akustičkog modela govora, ovo je zadnji korak pri kojem možemo procijeniti koliko je naučeni model dobro prilagođen slučaju za koji je ga koristimo. Osim navedenih uporaba, a s obzirom na prije opisanu evaluacijsku uporabu, koristi se i kada treba izabrati između više modela.

3.1.2 Vjerojatnost prikrivenih stanja za uočenu sekvencu

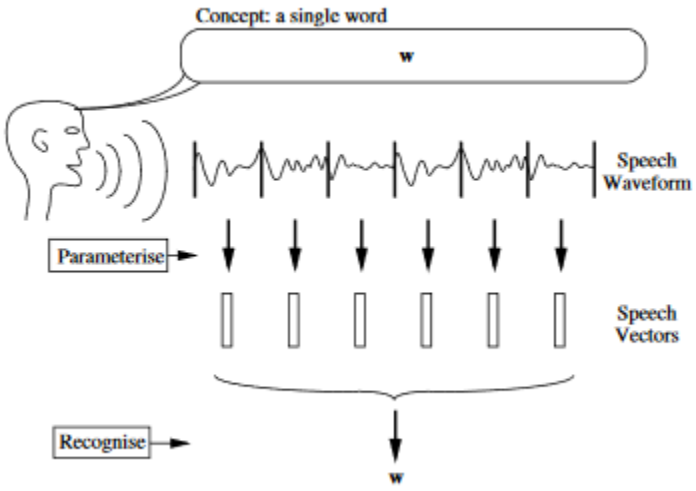
Za uočenu sekvencu $O = O_1{*}O_2{*}...{*}O_n$ treba izabrati optimalan redoslijed događaja $Q = q_1{*}q_2{*}...{*}q_n$ koji suvislo opisuje opservacije. Kod uporabe na kontinuiranom signalu govora rješavanjem ovoga problema možemo pronaći optimalan redoslijed stanja ili dobiti statističke podatke o pojedinom stanju.

3.1.3 Određivanje parametara modela

Određivanje parametara modela $\lambda = (A, B, \pi)$ kako bismo maksimizirali $P(O|\lambda)$ je u suštini optimizacija modela. Uočenu sekvencu koju koristimo za namještavanje parametara modela jos zovemo i trening sekvenca.

3.2 Primjena pri modeliranju signala govora

Signal govora određene riječi prikazan je kao vremenski redoslijed kodiranih spektralnih vektora uz pretpostavku da je signal kodiran koristeći knjigu kodnih riječi s M jedinstvenih spektralnih vektora. Prvi zadatak je izrada individualnih modela riječi određivanjem parametara modela.
Zatim se ispituje vjerojatnost prikrivenih stanja segmentacijom sekvenci riječi za treniranje u stanja te se proučavaju svojstva spektralnih vektora koji su doveli do opažanja za svako stanje. Cilj ovog koraka je poboljšanje parametra za bolje modeliranje redoslijeda izgovorenih riječi.
Kada je dobiveno nekoliko modela, testiraju se vjerojantosti uočenih sekvenci te izabire model koji najbolje opisuje slučajeve dane testiranjem. Jedna od metoda odabira modela je metoda najveće vjerojatnosti.
Za razliku od diskretnih, signal govora je kontinuiran.

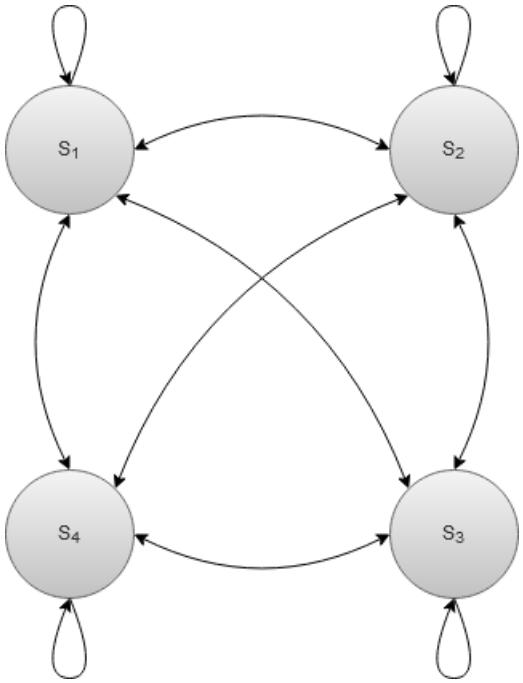


Slika 3: Modeliranje signala govora

3.3 Tipovi prikrivenih Markovljevih modela za signal govora

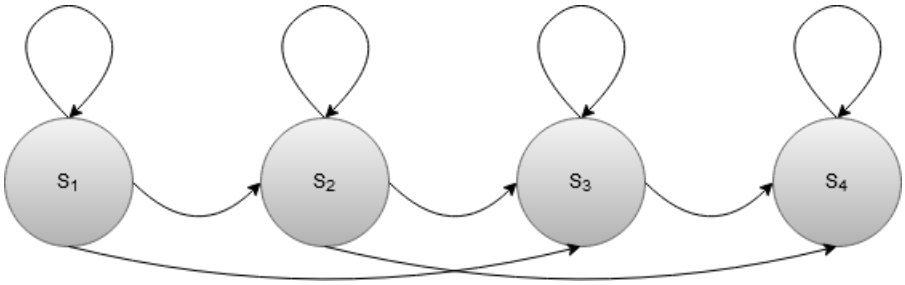
Stanja u prikrivenom Markovljevom modelu mogu biti povezana na mnogo različitih načina. Ergodički model podrazumijeva povezanost svih stanja tj. mogućnost prelaska u svako stanje iz svakog stanja. Takav tip povezanosti nije dostatan za modeliranja signala govora.

Pri modeliranju signala govora razumno je da povratak u prethodna stanja ne bi imao smisla, kako je svako novo opažanje (do kojega su doveli spektralni vektori) ujedno i novo stanje. Prethodna stanja odgovaraju prijašnjim opažanjima, pa model koji dopušta povratak u prijašnje stanje nije prikladan za opisivanje signala govora.



Slika 4: Prikriveni Markovljev model - ergodički tip

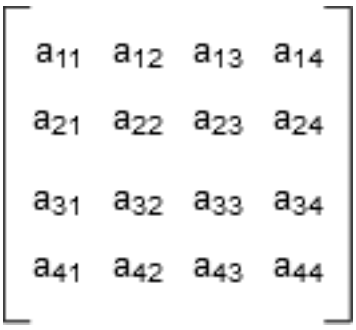
Osim nepovratnosti u prijašnja stanja, kod modeliranja signala govora osim monofona (jedne cjeline) imamo i trifone koji su cjelina sastavljena od tri monofona, kako i sama riječ naznačuje. Ovisno o bazi govora s kojom raspolažemo za treniranje modela, imamo primjere riječi i osim statistike koje riječi dolaze u parovima (u smislenom govoru postoje pravila i učestalosti pojavljivanja riječi prije ili nakon određene riječi) tako i unutar riječi postoji učestalost pojavljivanja fonema. Svaki fonem u riječi, osim prvoga i zadnjega, ima svojeg prethodnika i sljedbenika s kojim tvori cjelinu. Kako bismo u obzir uzeli i cjeline trifona, osim vjerojatnosti iduceg stanja (u trenutku t+1) treba uzeti i vjerojatnost stanja u trenutku t+2 u odnosu na trenutno stanje.



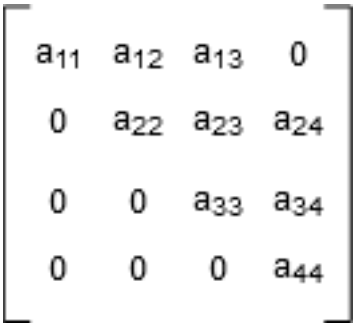
Slika 5: Prikriveni Markovljev model - primjer vrste za modeliranje signala govora

Iznad opisane vrste prikrivenih Markovljevih modela nisu jedine koje opisuju potencijalne modele, ali kako vrsta modela lijevo-desno (engl. left-right model) dobro opisuje signal govora zadržati ćemo se na usporedbi ergodičkog modela te lijevo-desno modela.

Nakon što smo definirali vjerojatnosti prijelaza između stanja za lijevo-desno model, bitno je i napomenuti da kad modeliramo signal govora moramo započeti u stanju 1 (S_1). To znači da



Slika 6: Vjerojatnosti prijelaza između stanja - ergodički model



Slika 7: Vjerojatnosti prijelaza između stanja - lijevo-desno model

vjerojatnosti da je prvo stanje u kojem se nalazimo različito od stanja 1 moraju biti jednaka nuli, a vjerojatnost da započinjemo u prvom stanju mora biti jednaka 1.

$$\pi_i = \begin{cases} 0, i = 11, i = 1 \end{cases}$$

4 Priprema podataka i izrada modela

Prije izrade akustičkog modela potrebno je pripremiti podatke koji će služiti za treniranje modela. Sustav za raspoznavanje govora sastoji se od jezičnog modela (ili gramatičke datoteke), akustičnog modela te dekodera.

- Jezični model sadrži riječi i vjerojatnosti da se pojave u sekvenci. Gramatička datoteka je manja od jezičnog modela te sadrži predefinirane kombinacije riječi.
- Akustički model sadrži statističku reprezentaciju različitih zvukova od kojih se jezični model ili gramatička datoteka sastoje.
- Dekoder je računalni program (engl. software) koji za izgovorene riječi pretražuje akustički model kako bi za njih pronašao odgovarajuće zvukove.

4.1 Gramatička datoteka u Juliusu

Kod Juliusa je gramatika potrebna za raspoznavanje podijeljena u .grammar (pravila za riječi koje bi sustav za raspoznavanje govora trebao raspoznati) i .voca (podjela riječi na kategorije skupa s fonemima koji tvore izgovorenu riječ) datoteke.

4.2 .grammar datoteka

Pravila za definiranje dozvoljenih riječi definiranih u ovoj datoteci odgovaraju prerađenom BNF formatu. BFN je skraćenica od Backus-Natur form i služi za definiranje formata u kojem će potrebni podaci biti unošeni.

```
1 S: NS_B SENT NS_E
2 SENT: CALL_V NAME_N
3 SENT: DIAL_V DIGIT
```

Slika 8: Primjer gramatičke datoteke formatirane za Julius

Na slici 7 prikazana je definicija jednostavne gramatičke datoteke. Simboli u .grammar datoteci dijele se na terminalne i neterminalne. "S" je simbol koji označava početnu rečenicu (govorni iskaz), a "NS_B" i "NS_E" označavaju tišinu koja je očekivana prije i nakon govornog iskaza (engl. utterance) te moraju biti definirani kod definiranja rečenice. Oni predstavljaju terminalne simbole što znači da su konstantne vrijednosti koje moraju biti definirane u gramatičkoj datoteci.
"SENT" je neterminalni simbol tj. simbol koji možemo proizvoljno unijeti te je ujedno (u slučaju prikazanom na slici 7) i simbol koji označava početnu rečenicu. U donja dva reda gramatičke datoteke imamo simbol početne rečenice zamijenjen s terminalnim simbolima. "CALL_V" "NAME_N" i "DIAL_V" "DIGIT" su terminali koji zamjenjuju početnu rečenicu te također moraju biti definirani u .voca datoteci.

4.3 .voca datoteka

Svi terminalni simboli moraju biti daljnje definirani u .voca datoteci.

Kao što je i vidljivo na slici 8, za terminalne simbole koji opisuju početnu rečenicu u .voca datoteci naznačimo riječi koje se mogu nalaziti na njihovom mjestu skupa sa fonemima koji tvore te riječi.
Nakon što su .grammar i .voca datoteke gotove, koristeći skriptu tvore se .dfa i .term datoteke koje sadrže informacije o konačnom automatu i .dict datoteka koja sadrži informacije o rječniku. Sve tri datoteke imaju isto ime te su u Julius formatu.

4.4 Rječnik izgovora (engl. pronunciation dictionary)

HTK treba minimalno trideset do četrdeset rečenica koje sadrže između osam i deset riječi u rječniku izgovora kako bi preveo snimke govora i transkripcije u akustični model. S obzirom na manjak rečenica i riječi u .grammar i .voca datotekama dodaju se rečenice kako bi HTK mogao obaviti prevođenje.
Za stvaranje rječnika izgovora potreban je popis svih korištenih riječi. Naredba za stvaranje rječnika izgovora je HDMan kojom dobijemo listu korištenih fonova te rječnik izgovora.

5 Rezultati

Rezultati.

```
% NS_B
<s> sil

%NS_E
</s> sil

%CALL_V
PHONE f ow n
CALL k ao l

%DIAL_V
DIAL d ay l

%NAME_N
STEVE s t iy v
YOUNG y ah ng

%DIGIT
FIVE        f ay v
FOUR        f ow r
NINE        n ay n
EIGHT       ey t
OH          ow
ONE         w ah n
SEVEN       s eh v ih n
SIX         s ih k s
THREE       th r iy
TWO         t uw
ZERO        z iy r ow
```

Slika 9: Primjer .voca datoteke formatirane za Julius

```
ABALON          [ABALON]        ae b ah l ow n sp
ABDOMINALS      [ABDOMINALS]    ae b d aa m ah n ah l z sp
ABOLISH         [ABOLISH]       ah b aa l ih sh sp
ABOUNDING       [ABOUNDING]     ah b aw n d ih ng sp
ABOUT           [ABOUT]         ah b aw t sp
ACCOUNT         [ACCOUNT]       ah k aw n t sp
ACHIEVE         [ACHIEVE]       ah ch iy v sp
ACTUAL          [ACTUAL]        ae k ch ah w ah l sp
ACUPUNCTURE     [ACUPUNCTURE]   ae k y uw p ah ng k ch er sp
ACUTE           [ACUTE]         ah k y uw t sp
```

Slika 10: Dio rječnika izgovora

6 Zaključak

Prateći uputstva izrađen je akustički model za raspoznavanje jednostavnih riječi ili točnije rečeno naredbi. Broj korištenih riječi i primjera u modelu je nedostatan za izradu modela visoke preciznosti. Zbog minimalnog broja rečenica i riječi u njima, od kojih je većina dodana samo kako bi HTK mogao prevesti snimke govora i njihove transkripcije u rječnik izgovora očekivano je da preciznost modela neće biti velika. Kod tako malih baza postoji mogućnost da neki fon uopće nije zastupljen, a u ovom slučaju je problem što nekolicina fonova nemaju minimalnih 5 pojavljivanja. Osim toga, problem je i u korištenoj opremi pri snimanju uzoraka kao i za testiranju.

Literatura

[1] Lawrence R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*

[2] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, *The HTK Book (for HTK Version 3.3)*

[3] Nicolas Moreau, *HTK (v.3.1): Basic Tutorial*

[4] Akinobu Lee , Tatsuya Kawahara , Kiyohiro Shikano, *Julius — an Open Source Real-Time Large Vocabulary Recognition Engine*

[5] *Julius*, dostupno na `https://julius.osdn.jp/en_index.php`.

[6] *Voxforge tutorial*, dostupno na `http://www.voxforge.org/home/dev/acousticmodels/windows/create/ht`